Environmental Health Perspectives
Vol. 96, pp. 113–120, 1991

Analysis and Comparison of Information and Data Recorded in Carcinogenicity and Genotoxicity Databases

by Paolo Romano,* Ottavia Aresu,† Barbara Parodi,† Davide Malacarne,* Gianfranco Castagneto,‡ Silvio Parodi,*† and Tiziana Ruzzon†

The Interlab Project is a university–industry joint project recently funded by the Italian government as part of the improvement of the Italian research infrastructure; among its short-term goals are the implementation of data banks of biomedical interest and the spread of informatic tools for biomedical research. Results of both long-term assays of carcinogenicity in rodents and short-term *in vitro* and *in vivo* tests of genotoxicity are relevant for a wide body of users, ranging from carcinogenesis research laboratories to industries and governmental agencies. To evaluate the most appropriate ways of spreading information on these experiments, a detailed analysis on information recorded in available databases has been carried out. Furthermore, the contents of the most known databases have been compared, with respect to a specific compound, to evaluate both the overall reliability of these systems, compared to longer and more complex assessments carried out manually starting from bibliographic searches, and the level of concordance among them.

Introduction

Telematics, the synergic use of informatics and communication systems to improve the transmission and sharing of data among computers at an international level, has had an enormous impact on the research environment. In recent years, a number of networks have been set up, and many on-line data banks have been created. Toxicity research, in particular that relating to carcinogenicity and genotoxicity, have become involved and somehow benefited from these initiatives. Moreover, the dramatic improvement of electronic technologies that has led to the design of high-performance, low-cost computers, and the sharpening of software methodologies, which in turn has led to the development of standardized database management systems, has given rise to the establishment of many databases on the same topics.

Thus, at the moment, a number of different sources for carcinogenicity and genotoxicity data is available. If the former are relatively few, this is not the case for the latter, since many new results are continuously being published. Availability of these tools to the end users has been guaranteed by means of a number of systems, ranging from literature reports to personal computer software, on-line data banks and CD-ROM (compact disk-read-only memory).

*Institute of Clinical and Experimental Oncology, University of Genoa, viale Benedetto XV 10, 16132 Genova, Italy.
†National Institute for Cancer Research, viale Benedetto XV 10, 16132 Genova, Italy.
‡Ansaldo SpA, via Pieragostini, 50, 16151 Genova, Italy.

Due to the accelerated development of the state of the art in data management systems and to the high economic investments needed to constantly maintain an advanced position in this field, the most important international regulatory agencies and cancer institutes do not always become involved in creating these systems. There is a need for a single, comprehensive, exhaustive database, easily accessible to a wide body of users. To highlight the problems that must be solved to achieve this goal, we examine the current situation in terms of databases available, information taken into account, and overlapping of data from the point of view of the end user.

Interlab Project

The Interlab Project is a university–industry joint project (*1*). The promoting institutions for the project are the National Institute for Cancer Research (IST) of Genoa; the InterUniversity Center for Cancer Research (CIRC), grouping five Italian universities; and Ansaldo SpA, an Italian leader in information systems. The project was funded on a 2-year scheme in June 1989 by the Italian Ministry for University and Scientific and Technological Research, with the goal of improving the Italian research infrastructure.

The main objectives of Interlab are to improve existing collaborative links among biomedical research centers operating in kindred fields and to spread the use of computer tools devoted to this research area in Italy, where background in informatics and awareness of its relevance are still lacking. A communication network has been set up to allow for easy and steady communica-

tion among institutes, whose researchers can exchange messages by means of an electronic mail system. It is planned that, in the future, a bulletin board system will also be available, as well as a custom service for bibliographic searches.

Furthermore, centralized, on-line factual data banks on availability of biological material in Italian laboratories have been implemented to allow for quick, exhaustive, and easy retrievals of constantly updated information in research areas in which few data were available. Personal computer versions of these databases are being created to help researchers maintain their collections of biological materials and guarantee a steady flow of up-to-date information.

Among the main design criteria, two are particularly worth mentioning because they give the system its specificity. These criteria are the use of the relational approach in defining data structures and the particular care in designing a friendly user interface. The relational approach has been adopted instead of the more traditional information retrieval approach in consideration of the nature of the information to be recorded.

Apart from its theoretical basis, the most evident characteristic of the data bank is that almost all data are coded and recorded in structured fields. This leads to the creation of complex data structures by means of which "normal data formats," i.e., formats in which data are recorded without any redundancies, can be obtained. One of the advantages of this approach is that searches can only be carried out in specific contexts, i.e., with respect to a specific information, thus both avoiding confusion arising from coincidental correspondence of terms and guaranteeing their exhaustiveness. Furthermore, searches are executed in a very efficient way and can thus be performed on almost all kinds of computers, independent of their speed and capacity. An automatic validation of data being inserted can be carried out by the system as well.

Moreover, since the relational approach leads to specific data structures, any ad hoc queries, specifying which terms can be searched and in which context, must be defined, and ad hoc applicative software, aimed to the creation of the user interface, must be developed.

The databases have been implemented on a Unix-based microcomputer, and the relational database management system Oracle, a commercial software available worldwide, has been adopted. Essentially, Oracle has been chosen for its wide spectrum of versions, ranging from personal computers to mainframes, and for its good modularity and portability that make the creation of database versions for other computers and of new releases easier and quicker.

The user interface is always a fundamental part of the system because it determines the real accessibility of the data. In our case, it also had to be easy to use and as clear as possible for people without specific skills in informatics. It has been carefully designed for general features, which are valid for all the applications, and for specific features, which are valid only in specific contexts, such as insertion and query.

Apart from having masked the Unix operating system and SQL (Structured Query Language, the standard query language for relational systems) to the end users, extensive use of menus and of contextual helps has been made. Furthermore, to simplify the interaction between the user and the applications, a limited number of function keys is used and, when possible, the words taken from the informatics jargon, like field, record, and block, have been substituted with more widely used terms.

Among user interface specific features, particularly relevant are those devoted to the optimization of the use of controlled vocabularies, such as the extensive use of mnemonic codes instead of the complete terms and an automatic display of the list of the available items. Moreover, data are validated during insertion, queries are defined according to high-level macroinformation, and data are presented in coherent subsets.

Until now, three databases have been implemented and are available on-line. The first one relates to cell lines (CLDB), the second to HLA-typed B-lymphoblastoid cell lines (BLDB), and the third to oligonucleotides (MPDB). CLDB contains data on 720 cell lines available in Italian laboratories. More than the 60% of these lines are original, that is, not described in any other commercial or scientific catalog. In fact, CLDB data collection highlighted the presence of many well-characterized, small collections of cell lines. CLDB data structure is quite complex and is based on two substructures. The first relates to information that univocally identifies the cell line, the second to information that is specific for a laboratory in which the cell line is collected. Among the former are the name, the origin (species, strain, sex, etc.) and possible transformations; among the latter, it considers culture conditions and validation assays performed. Controlled vocabularies have been defined for most information. Among them are species and relative strains, morphologies, tumors, transforming agents, applications, and functions.

Searches can be carried out using three different approaches: by name, by origin, and by function. Using the first approach, the search can be conditioned on the basis of the name, the presence in a given catalog and/or the identification code in a catalog. The query by origin can be used to retrieve cell lines having given species, strain, tissue, tumor, and pathology. Finally, the query by function relates to cell lines applications and specific functions. A new approach, based on a query related to the transforming agent, will be added in the near future. Following the retrieval of desired cell lines, information can be displayed according to coherent subsets, which are identification, origin, specification, ownership and culture data. Both detailed and synthetic reports can be generated with reference to one single cell line or many cell lines having some common characteristics.

BLDB and MPDB have been designed using the same criteria. BLDB contains data on approximately 750 B-lymphoblastoid lines available from the laboratories of the European Collection for Biomedical Research (Essen, Germany and Genoa, Italy). Data from two other European collections are being added. At the moment, the prototype of MPDB contains data on oligonucleotides produced by the internal service of the National Institute for Cancer Research of Genoa. It will be flanked by a service for the production of custom oligonucleotides.

Carcinogenicity and Genotoxicity Databases

There are many carcinogenicity and/or genotoxicity databases that are available to the end users. Some of them are hosted by computers of large information companies and can be searched on line. Others are not available on line, but their complete, up-to-date dump can be obtained from database administrators on floppy disks or tapes, at times with some ad hoc software, that allows for their management. Finally, some are available to the end users only in a printed format.

Often, when biologists want to search these databases, they are not fully aware of the main goals and specificities of each of them. The databases included in this analysis have been chosen on the basis of their availability and relevance, relevance measured in terms of quantity of data, promoting institution, international agencies involvement and geographic origin, and have been analyzed from the point of view of these unpracticed biologists.

In regard to on-line databases, Registry of Toxic Effects of Chemical Substances (RTECS) (*2*), Chemical Carcinogenesis Research Information System (CCRIS) (*3*), and Environmental Chemical Data and Information Network (ECDIN) (*3*) have been considered. RTECS is a factual, nonbibliographic data bank, built and maintained by the National Institute for Occupational Safety and Health (NIOSH) and hosted by a number of well-known host computers in the United States, Europe, and Australia; it contains mutagenesis studies on nearly 10,000 substances and carcinogenesis studies on about 3,400 substances. CCRIS is hosted by Chemical Information Systems, Inc. (CIS) and the National Library of Medicine (NLM); it contains carcinogenicity, co-carcinogenicity, and mutagenicity data on 1,269 substances. ECDIN is a factual data bank, created by the Joint Research Centre (JRC) of the Commission of the European Communities (CEC) at Ispra, Italy; it is hosted by Datacentralen.

Even if these data banks have different objectives, they are all comprehensive in the sense that they present both carcinogenicity and genotoxicity data. In addition to these, another comprehensive database has been included in the analysis, the Biological Database (BL-DB) (*4*), although it is not available on line. It is a fact database, containing data on mutagenicity and carcinogenicity.

Databases that are specific for carcinogenicity or genotoxicity have also been included in the analysis. They are, respectively, Carcinogenic Potency Database (CPDB) (*5*) and the Gene-Tox Carcinogen Database (GTCDB) (*6*) for carcinogenicity and the Genetic Activity Profile Database (GAP) (*7,8*) and the GEN Database (GEN) (*9*) for genotoxicity. CPDB contains standardized data on 4000 animal experiments with about 1000 chemical compounds. GTCDB contains data on more than 500 selected chemicals. GAP provides activity profiles and corresponding listings of data and references for each chemical analyzed. Two data sets are included in the GAP software: one related to the International Agency for Research on Cancer (IARC) (277 agents) and one related to the U.S. Environmental Protection Agency (EPA) (167 agents).

Selection of data for the analysis has been carried out manually for CPDB and GTCDB and on the basis of ad hoc management software distributed with the databases for GAP and GEN.

Objectives and Results

The research was mainly aimed at *a*) comparing the types of information recorded in each database, *b*) determining a common, basic data set, *c*) evaluating data overlapping between databases, *d*) evaluating general agreement/disagreement among results reported by different databases, and *e*) evaluating general reliability of databases, as tools able to provide the basis for rapid and efficient synthetic evaluations on a given chemical or groups of chemicals.

A detailed analysis on information recorded in available carcinogenicity databases has been carried out. Data have been subdivided into two groups, devoted, respectively, to the description of the compound and the experiments, and each of these into many coherent subgroups (Tables 1 and 2).

As far as the description of the compound (Table I) is concerned, although identification data seem adequate for all the databases, the other subgroups of information are lacking. In particular, other physicochemical parameters that could be relevant for carcinogenicity, such as hydrophobicity and various types of structural alerts, and pharmacokinetics data are not reported at all. Furthermore, recognized evaluations, such as IARC and National Cancer Institute/National Toxicology Program (NCI/NTP) classifications, are generally neither listed nor sufficiently highlighted. Supplementary information, which could be relevant for accessing compound data on the basis of the chemical class or use, are reported very rarely. Finally, links to other databases are almost completely absent.

In regard to experiment description (Table 2), the set of information taken into account is almost the same for all the databases, but, with the exception of species, strains and sex of the animals and route of administration of the compound, they are described in a number of different, nonstandardized ways. This is particularly evident for information related to experimental design and results. Other information on results, such as tumor latency, which is relevant for risk assessment, are normally absent. Quantitative evaluations are rarely present and bibliographic references are not standardized.

Data overlapping has been evaluated by comparing citations and single long-term animal experiments reported by each carcinogenicity database. To this end, a specific compound (benzene) has been chosen and all related information has been selected from the databases and analyzed (Tables 3–5). Every carcinogenicity study singly identifiable on the basis of bibliographic reference, species, strain, sex and route of administration has been considered as a separate experiment. Data show that NCI/NTP experiments are normally listed, even if not all the experiments that were carried out in this context are reported (Table 3). Some misunderstanding can arise in regard to IARC monographs, which being surveys, do not list any original experiment: in two cases a monograph was reported as an original reference for the experiment, while, in the others, references to original experiments reported also in an IARC monograph were given. Apart from NCI/NTP technical reports and IARC monographs, databases reported, in most cases, a great number of original experiments (ranging from 7 to 12), but the overlap was poor: only 5 out of 37 experiments were reported in more than one database. Total experiment redundancy (still excluding those of the NCI/NTP and IARC), corresponding to the percentage of experiments reported more than once, is thus about 14%.

A similar situation can be shown by analyzing citations only (Table 4). In this case, since ambiguities possibly arising from different interpretation of results shown in papers are absent, data are more readable. Total redundancy, corresponding to the percentage of references cited in more than one database, is ca. 22%. Citations and citing databases are reported in Table 5.

The low redundancy that has been found can be explained on the basis of many different reasons. One possible explanation is that more than one paper can present and discuss the same original experiment, possibly with some marginal updating or deeper analysis. In this case, the same data could be inserted in

Table 1. Compound identification data in carcinogenicity databases.

	Database[a]						
	CPDB	CCRIS	RTECS	GTCDB	BL-DB	ECDIN	HypDB
Identification							
Name	X	X	X	X	X	X	X
Synonyms and trade names	X		X		X	X	X
CAS registry number	X	X	X	X	X	X	X
Other international reference numbers and names					X	X	X
Chemical properties							
Chemical formula			X		X	X	X
Formula mass			X				
Chemical structure			X[b]		X[b]	X[b]	X
Other physicochemical parameters					X[c]	X[d]	X[e]
Pharmacokinetics							X
Overall evaluation							
Own evaluation			X	X	X		X
IARC evaluation							X
NCI/NTP evaluation			X				X
Other evaluations					X[f]		X
Supplementary information							
Chemical class					X		X
Major uses		X	X		X		X
Other key words					X		X

Abbreviations: CPDB, Carcinogenic Potency Database; CCRIS, Chemical Carcinogenesis Research Information System; RTECS, Registry of Toxic Effects of Chemical Substances; GTCDB, Gene-Tox Carcinogen Database; BL-DB, biological database; ECDIN, Environmental Chemical Data and Information Network; HypDB, hypothetical database.

[a]CCRIS and RTECS are on-line data banks. An X indicates information taken into account by the databases.

[b]Wiswesser line notation.

[c]Melting and boiling points.

[d]About 20 different parameters.

[e]Examples are ionic status, structural alerts, hydrophilicity/phobicity.

[f]RTECS toxic dose.

Table 2. Experiment description data in carcinogenicity databases.

	Database[a]						
	CPDB	CCRIS	RTECS	GTCDB	BL-DB	ECDIN	HypDB
Animals							
Species	X	X	X	X	X	X[b]	X
Strain	X	X	X	X	X	X[b]	X
Sex	X	X	X	X	X	X	X
Other information					X[c]		
Experimental design							
Route	X	X	X	X	X	X	X
Doses	X	X[d]	X[e]		X	X[f]	X
Duration	X	X[d]	X[e]	X	X	X[g]	X
Sample size	X				X		X
Compound's purity					X		
Results							
Target organ	X	X[h]	X[i]	X[h]	X[j]	X[h]	X
Tumor type	X	X[h]	X[i]	X[h]	X[j]	X[h]	X
Tumor incidence	X				X[j]	X	X
Tumor latency						X	X
Evaluation							
Qualitative	X[k]	X	X[i]	X	X		X
Quantitative	X		X[e]		X		X
Bibliography	X	X	X	X	X	X	X

Abbreviations: CPDB, Carcinogenic Potency Database; CCRIS, Chemical Carcinogenesis Research Information System; RTECS, Registry of Toxic Effects of Chemical Substances; GTCDB, Gene-Tox Carcinogen Database; BL-DB, Biological Database; ECDIN, Environmental Chemical Data and Information Network; HypDB, hypothetical database.

[a]CCRIS and RTECS are on-line data banks.

[b]Free-text description of species and strain.

[c]Age, weight.

[d]Free-text description of nonstandardized doses and duration.

[e]Free-text description of duration and either lowest dosage inducing a significant increase in tumor incidence or dosage inducing a significant increase in tumor incidence.

[f]Separate description of single dose, total dose, and comment on dose.

[g]Separate description of frequency and duration of administration.

[h]Free-text description of target organ and tumor.

[i]Free-text description of target organ, tumor, and effects.

[j]Tabular format and short text description.

[k]Author's opinion, if stated.

Table 3. Reported experiments and overlap in carcinogenicity databases: benzene (CAS no. 71-43-2).

	NTP[a]	IARC[b]	Total (uniques)[c]	Overlap[d] RTECS	CCRIS	CPDB	GTCDB	ECDIN
RTECS	1	3[e]	12 (7)	—	0	2	2	1
CCRIS	4	3	1 (1)	0	—	0	0	0
CPDB	4	4[e]	10 (8)	2	0	—	0	0
GTCDB	4	2	12(10)	2	0	0	—	0
ECDIN	0	6[e]	7 (6)	1	0	0	0	—
Reported experiments: 37		Redundancy: 5/37 (~ 14%)						

Abbreviations: CPDB, Carcinogenic Potency Database; CCRIS, Chemical Carcinogenesis Research Information System (on-line data bank); RTECS, Registry of Toxic Effects of Chemical Substances (on-line data bank); GTCDB, Gene-Tox Carcinogen Database; ECDIN, Environmental Chemical Data and Informative Network (on-line data bank).

[a]Number of experiments reported from the NCI/NTP Technical Report.

[b]Number of experiments reported from *IARC Monographs* as an original source of data.

[c]Total number of experiments and, in parentheses, the number of experiments that are not reported in any other database, excluding NCI/NTP Technical Report and *IARC Monographs*.

[d]Number of experiments reported by each couple of databases.

[e]Number of references cited both by the database and the *IARC Monograph*.

Table 4. Citations reported and overlap in carcinogenicity databases: benzene (CAS no. 71-43-2).

	NTP[a]	IARC[b]	Total (uniques)[c]	Overlap[d] RTECS	CCRIS	CPDB	GTCDB	ECDIN
RTECS	Yes	2[e]	9(5)	—	0	1	2	1
CCRIS	Yes	Yes	1(1)	0	—	0	0	0
CPDB	Yes	2[e]	4(3)	1	0	—	0	0
GTCDB	Yes	Yes	4(2)	2	0	0	—	0
ECDIN	No	3[e]	4(3)	1	0	0	0	—
Reported citations: 18		Redundancy: 4/18 (~ 22%)						

Abbreviations: CPDB, Carcinogenic Potency Database; CCRIS, Chemical Carcinogenesis Research Information System (on-line data bank); RTECS, Registry of Toxic Effects of Chemical Substances (on-line data bank); GTCDB, Gene-Tox Carcinogen Database; ECDIN, Environmental Chemical Data and Informative Network (on-line data bank).

[a]Citation of NCI/NTP Technical Report.

[b]Citation of *IARC Monographs*. "Yes" indicates that the monographs were used directly as a source of information.

[c]Total number of citations and, in parentheses, the number of citations that are not reported in any other database, excluding NCI/NTP Technical Report and *IARC Monographs*.

[d]Number of citations reported by each couple of databases.

[e]Number of references cited both by the database and the *IARC Monograph*.

Table 5. Citations reported for benzene (CAS no. 17-43-2) in carcinogenicity databases.

Reference	Database
Maltoni et al. (*10*)	RTECS, GTCDB
Maltoni et al. (*11*)	CPDB
Baldwin et al. (*12*)	RTECS
Snyder et al. (*13*)	GTCDB
Kirschbaum et al. (*14*)	ECDIN, IARC
Hiraki et al. (*15*)	ECDIN
Cronkite et al. (*16*)	RTECS
Snyder et al. (*17*)	CPDB
Lignac (*18*)	RTECS
Lignac (*19*)	RTECS, ECDIN, IARC
Maltoni and Scarnato (*20*)	RTECS, CPDB, IARC
Maltoni et al. (*21*)	CCRIS
Sellakumar et al. (*22*)	RTECS
Amiel (*23*)	ECDIN, IARC
Snyder et al. (*24*)	CPDB, IARC
Cronkite et al. (*25*)	RTECS, GTCDB
Stoner et al. (*26*)	RTECS
Goldstein et al. (*27*)	GTCDB

Abbreviations: IARC, International Agency for Research on Cancer; CPDB, Carcinogenic Potency Database; CCRIS, Chemical Carcinogenesis Research Information System; RTECS, Registry of Toxic Effects of Chemical Substances; GTCDB, Gene-Tox Carcinogen Database; ECDIN, Environmental Chemical Data and Information Network.

different databases with different references. Another explanation is that different insertion criteria can lead to different selection of original works. Finally, because bibliographic databases can be geographically biased, the use of different host computers could hide works published on secondary journals.

Though these reasons can help to understand the reasons for the differences that we have pointed out, it should nonetheless be taken into account that the main goal of factual databases is to give end users sufficient data without reading the original papers. From this point of view, the current situation could be misleading and could give the impression that there are more data than in reality and could lead to overestimating experiments reported more than once.

In regard to the evaluation of agreement among results reported by different databases, genotoxicity databases have been compared, still with respect to benzene; this compound shows a somewhat puzzling behavior (Tables 6–9). Even if, considering the whole set of experiments reported by the IARC dataset of the GAP database, the compound should be considered as prevailingly negative because it has only a 25% positive result rate (Table 6), a clear positiveness is shown for *in vivo* tests, here represented only by chromosomal damage assays. More specifically, relatively few *in vitro* DNA damage short-term experiments show a clearly negative pattern, with positive results constantly lower than 27%. Conversely, chromosomal damage experiments show a clear negative behavior for *in vitro* tests, both with and without metabolic activation, and a clearly positive one

Table 6. Genotoxicity tests for benzene (CAS no. 71-43-2) as reported by the Genetic Activity Profile database (IARC data set).[a]

	In vitro			*In vitro* with activation			*In vivo*			Total		
	−	+	%+	−	+	%+	−	+	%+	−	+	%+
DNA damage	8	3	27	5	0	0				13	3	19
Chromosomal damage	28	7	20	16	7	30	4	23	85	48	37	44
Mutation	47	9	16	47	4	8				94	13	12
Total	83	19	19	68	11	14	4	23	85	155	53	25

[a]For each end point, the total number of positive and negative results listed in the Genetic Activity Profile database (GAP) is reported, together with the percentage of positive results. Row and column totals are reported. Transformation assays and two other nonclassifiable assays have not been considered. Relatively few weak responses, both positive and negative, have been included. Inconclusive results have been discarded.

Table 7. Comparison among genotoxicity databases: benzene (CAS no. 71-43-2).[a]

	In vitro			*In vitro* with activation			*In vivo*			Total		
	−	+	%+	−	+	%+	−	+	%+	−	+	%+
DNA damage												
GAP	8	3	27	5	0	0				13	3	19
ECDIN	10	2	17	1	0	0				11	2	15
Chromosomal damage												
GAP	28	7	20	16	7	30	4	23	85	48	37	44
ECDIN	1	0	0				2	12	86	3	12	80
Mutation												
GAP	47	9	16	47	4	8				94	13	12
ECDIN				9	0	0	1	0	0	10	0	0
CCRIS	3	0	0	5	0	0				8	0	0

Abbreviations: GAP, Genetic Activity Profile database; ECDIN, Environmental Chemical Data and Information Network; CCRIS, Chemical Carcinogenesis Research Information System.

[a]For each end point, the total number of positive and negative results are reported together with the percentage of positive results. Row totals are also reported. For GAP data, see Table 6. No DNA damage and chromosomal damage data were available on CCRIS. RTECS was not considered because it does not report effects on single genotoxicity tests.

Table 8. Comparison among genotoxicity databases: benzene (CAS no. 71-43-2).[a]

	In vitro			*In vivo*		
	−	I	+	−	I	+
Chromosomal damage						
GAP	13	5	4	2	0	5
GEN	1	0	0	0	1	2
Mutation						
GAP	26	6	2			
GEN	2	1	2			

Abbreviations: GAP, Genetic Activity Profile database; GEN, GEN database.

[a]For each end point, total negative, inconclusive (I), and positive figures are reported. Each test system has been considered only once and if the same test was performed in more than one experiment, an overall evaluation is reported. For GAP data, the test has been considered negative when less than 25% of results were positive and otherwise positive. GEN data are already listed as overall evaluations in the database.

for *in vivo* tests; the percentage of positive results ranges from 20 to 30% in the former case and attains 85% in the second, with a general mean of 44%. Finally, mutation experiments show a clearly negative behavior, with percentages of positive results ranging from 14 to 19%.

Table 9. Comparison between Genetic Activity Profile (GAP) database and bibliographic searches (BS): benzene (CAS no. 71-43-2).[a]

	In vitro			*In vivo*			Total		
	−	+	%+	−	+	%+	−	+	%+
DNA damage									
GAP	13	3	19				13	3	19
BS	6	7	54	2	10	83	8	17	68
Chromosomal damage									
GAP	44	14	24	4	23	85	48	37	44
BS	16	16	50	11	83	88	27	99	79
Mutation									
GAP	47	9	16				47	9	16
BS	46	19	29	1	0	0	47	19	29
Total									
GAP	104	26	20	4	23	85	108	49	31
BS	68	42	38	14	93	87	82	135	62

[a]For each end point, the total number of negative and positive results listed in databases are reported, together with the percentage of positive results. Row and column totals are reported as well. For GAP data see Table 6. Data from bibliographic searches are courtesy of S. Grilli and A. M. Colacci of the Institute of Oncology of the University of Bologna, Italy.

These results are substantially confirmed by comparison with other databases (Tables 7 and 8), although fewer data are available. Although RTECS cannot be compared because it does not report results of single experiments, it is possible for ECDIN and CCRIS, and data available confirm benzene behavior for both mutation and chromosomal damage assays.

To compare GAP and GEN data, the former must be re-examined. In fact, instead of listing the results of all published experiments, GEN reports an overall evaluation of all experiments related to one specific assay. Re-examination of GAP data has been carried out considering a test negative when more than 75% of the results were negative, inconclusive when less than 50% of the results were positive, and positive otherwise. Even if much fewer data are available after this re-examination (Table 8), the previously shown behavior of benzene is substantially maintained and confirmed by both databases.

This substantial consistency of experiments on benzene reported on many different databases demonstrates that, although some of them are lacking for particular types of tests, end users can trust databases to simplify, improve, and speed up their work. The great differences existing among databases, nevertheless, indicate the necessity of considering some databases as more reliable than others.

The benzene activity profile in short-term genotoxicity assays, as provided by databases, has also been compared to the results of an extensive continuous bibliographic search (Table 9) that is being carried out by researchers of the Institute of Oncology of the University of Bologna (S. Grilli, personal communication). The result of bibliographic search (BS) and analysis lists 217 experiments on benzene versus the 157 reported in GAP, i.e., approximately 38% more. The distribution of these assays (Table 9) highlights that these are not extra references not yet included in GAP, but that the two sets of data are different. Indeed, for example, BS reports 94 *in vivo* chromosomal damage experiments and 32 *in vitro* chromosomal damage experiments, while GAP reports only 27 and 58, respectively. The overall ratio of positive experiments is higher for BS than for GAP (62% against 31%). This is only partly due to the presence of 10 clearly positive *in vivo* DNA damage experiments. In fact, there is a constant higher ratio of positive results in BS than in GAP for all groups of experiments. This does not modify existing differences between *in*

vitro and *in vivo* assays and between mutation and chromosomal damage assays. This comparison, however, seems to suggest that experiences and knowledge present in some institutions should not be missed and that a multinode data input scheme, in which every node is in charge of insertion of data relative to its main expertise, would be preferable if common criteria and standards of quality could be achieved.

Conclusions

Types of information recorded into six different carcinogenicity databases, three of which are already available on-line, have been compared to verify if a common format was used, thus allowing data interchange, and to identify a basic data set. This comparison showed an extremely diversified situation in which the physicochemical characterization of chemical compounds, as well as overall evaluations and inter-database references are poor. Furthermore, the description of the experiments were extremely variable and nonstandardized. Suggestions on information, not yet taken into account, but relevant in view of a unified data set for carcinogenicity, have been given. The comparison of experiments on benzene and of respective bibliographic references reported by carcinogenicity databases showed that each of them lists many original works not reported by any other database, thus producing a low redundancy of references. Explanations for this unexpected result have been proposed, though this diversified reality is actually what appears to end users.

Results reported on benzene by four different genotoxicity databases have been compared to verify the general agreement among them and their overall reliability. Results showed the same, well-known, global activity profile for all the databases, though only GAP seemed to present data on all different test systems.

Finally, GAP genotoxicity results for benzene have been compared with data obtained by means of a continuous bibliographic search. This comparison showed, on one hand, that GAP substantially presents a true image of reality and, on the other, that also for a good database, a great percentage of short-term experiments can still be missed.

In conclusion, this work shows that *a*) databases can be extremely useful for researchers in the fields of carcinogenicity and genotoxicity because they can unambiguously represent reality and prevent, at the same time, long and expensive surveys of the original data, *b*) a common basic data set for carcinogenicity and genotoxicity does not yet exist and data cannot be exchanged easily among databases, *c*) the availability of many databases does not help the end users, instead it can create misunderstanding, overestimation, or confusion, and *d*) an effort should be made to define a common reference format, identify, and support the best databases, even by multiplying the input nodes, to achieve database exhaustiveness.

This work has been partially funded by the Italian Ministry of University and of Scientific and Technological Research within the sphere of the Interlab Project. The authors thank Professor S. Grilli and Dr. A. M. Colacci of the Institute of Oncology of the University of Bologna, Italy, for their kindness in providing data, which they are continuously collecting from published literature, on short-term tests on benzene. The authors also thank Dr. M. Evangelisti and Dr. A. Bogliolo of the Scientific Information and Documentation Service of the Library of the National Institute for Cancer Research of Genoa for support in carrying out searches about on-line databases.

REFERENCES

1. Parodi, B., Romano, P., Aresu, O., Manniello, A., Vitiello, E., Iannotta, B., Ruzzon, T., and Santi, L. The Interlab Project: data bases for biomedical research. Chemistry Today 8: 23–25 (1990).
2. Sweet, D. V., Ed. Registry of Toxic Effects of Chemical Substances (RTECS), 1986 Ed. National Center for Occupational Safety and Health, Cincinnati, OH, 1987.
3. Guida alle basi dati 1990. Medianet ed., no. 2, Milano, Italy, 1990.
4. Hayashi, M., Nakadate, M., Osada, T., Ishibe, T., Tanaka, S., Maekawa, A., Sofuni, T., Nakata, Y., Kanoh, N., Hashiba, S., Takenaka, Y., and Ishidate, M., Jr. A fact database for toxicological data at the National Institute of Hygienic Sciences, Japan. Environ. Health Perspect. 96: 57–60 (1991).
5. Swirsky Gold, L., Sawyer, C. B., Magaw, R., Backmann, G. M., de Veciana, M., Levinson, R., Hooper, N. K., Havender, W. R., Bernstein, L., Peto, Pike, M. C., and Ames, B. N. A Carcinogenic Potency Database of the standardized results of animal biossays. Environ. Health Perspect. 58: 9–319 (1984).
6. Nesnow, S., Argus, M., Bergman, H., Chu, K., Frith, C., Helmes, T., McGaughy, R., Ray, V., Slaga, T. J., Tennant, R., and Weisburger, E. Chemical carcinogens. A review and analysis of the literature of selected chemicals and the establishment of the Gene-Tox Carcinogen Data Base. Mutat. Res. 185: 1–195 (1986).
7. Waters, M. D., Stack, H. F., Brady, A. L., Lohman, P. H. M., Haroun, L., and Vainio, H. Appendix 1. Activity profiles for genetic and related tests. In: *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans.* Genetic and Related Effects: An Updating of Selected IARC Monographs from Volumes 1 to 42, Supplement 6. International Agency for Research on Cancer, Lyon, 1987 pp. 687–696.
8. IARC. Monographs on the Evaluation of Carcinogenic Risks to Humans. Genetic and Related Effects: An Updating of Selected IARC Monographs from Volumes 1 to 42, Supplement 6. International Agency for Research on Cancer, Lyon, 1987.
9. Würgler, F. E. Predicting mammalian mutagenesis by submammalian assays: an application of database GEN. Environ. Health Perspect. 96: 37–39 (1991).
10. Maltoni, C., Conti, B., and Cotti, G. Benzene: a multipotential carcinogen. *Results of long-term bioassays performed at the Bologna Institute of Oncology.* Am. J. Ind. Med. 4: 589–630 (1983).
11. Maltoni, C., Conti, B., Cotti, G., and Belpoggi, F. Experimental studies on benzene carcinogenicity at the Bologna Institute of Oncology: current results and ongoing research. Am. J. Ind. Med. 7: 415–446 (1985).
12. Baldwin, R. W., Cunningham, G. J., Partridge, M. W., and Vipond, H. J. Studies on the carcinogenicity of tricycloquinazoline. The effects of substitution in the peripheral carbocyclic rings on carcinogenic activity. Br. J. Cancer 16: 275–282 (1962).
13. Snyder, C. A., Goldstein, B. D., Sellakumar, A., Bromberg, I., Laskin, S., and Albert, R. E. Toxicity of chronic benzene inhalation: CD-1 mice exposed to 300 ppm. Bull. Environ. Contam. Toxicol. 29: 385–391 (1982).
14. Kirschbaum, A., and Strong, L. C. Influence of carcinogens on the age incidence of leukemia in the high leukemia F strain of mice. Cancer Res. 2: 841–845 (1942).
15. Hiraki, K., Irino, S., and Miyoshi, I. Development of subcutaneous sarcomas in Swiss mice given repeated injections of benzene. Gann 54: 427–431 (1963).
16. Cronkite, E. P., Drew R. T., and Bullis, J. E. Benzene toxicity and how to approach the problem of chemical leukemogenesis. Immunol. Hematol. Res. Monogr. 3: 156–160 (1984).
17. Snyder, C. A., Goldstein, B. D., Sellakumar, A., Wolman, S. R., Bromberg, I., Erlichman, M. N., and Laskin, S. Hematotoxicity of inhaled benzene to Sprague-Dawley rats and AKR mice at 300 ppm. J. Toxicol. Environ. Health 4: 605–618 (1978).
18. Lignac, G. O. E. Die Benzolleukämie bei Menschen und Weissen Mäusen. Klin. Wochenscr. 12: 109–110 (1933).
19. Lignac, G. O. E. *Benzene leukaemia in humans and albino mice.* Krankheitsforschung 9: 403–453 (1932).
20. Maltoni, C., and Scarnato, C. First experimental demonstration of the carcinogenic effects of benzene; long-term bioassays on Sprague-Dawley rats by oral administration. Med. Lav. 70: 352–357 (1979).
21. Maltoni, C., Conti, B., and Scarnato, C. Squamous cell carcinomas of the oral cavity in Sprague-Dawley rats, following exposure to benzene by ingestion. First experimental demonstration. Med. Lav. 73: 441–445 (1982).
22. Sellakumar, A., Albert, R. E., and Snyder, C. Carcinogenicity of benzene. Proc. Am. Assoc. Cancer Res. 25: 75 (1984).
23. Amiel, J. L. Negative test for the induction of leukaemia in mice by benzene. Rev. Franc. Etud. Clin. Biol. 5: 198–199 (1960).

24. Snyder, C. A., Goldstein, B. D., Sellakumar, A. R., Bromberg, I., Laskin, S., and Albert, R. E. The inhalation toxicology of benzene: incidence of hematopoietic neoplasms and hematotoxicity in ARK/J and C57BL/6G mice. Toxicol. Appl. Pharmacol. 54: 323–331 (1980).
25. Cronkite, E. P., Bullis, J., Inoue, T., and Drew, R. T. Benzene inhalation produces leukemia in mice. Toxicol. Appl. Pharmacol. 75: 358–361 (1984).
26. Stoner, G. D., Conran, P. B., Greisiger, E. A., Stober, J., Morgan, K. T., and Pereira, M. A. Comparison of two routes of chemical administration on the lung adenoma response in strain A/J mice. Toxicol Appl. Pharmacol. 82: 19–31 (1986).
27. Goldstein, B. D., Snyder, C. A., Laskin, S., Bromberg, I., Albert, R. E., and Nelson, N. Myelogenous leukemia in rodents inhaling benzene. Toxicol. Lett. 13: 169–173 (1982).